

**PENDETEKSIAN BAHASA PADA TEKS
MENGUNAKAN METODE N-GRAM**

Tugas Akhir



Oleh

NANDA RIO PRATAMA

22 05 3834



**PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TEKNOLOGI INFORMASI
UNIVERSITAS KRISTEN DUTA WACANA
TAHUN 2011**

PENDETEKSIAN BAHASA PADA TEKS MENGUNAKAN METODE N-GRAM

Tugas Akhir



Diajukan kepada Fakultas Teknik Informatika
Universitas Kristen Duta Wacana
Sebagai salah satu syarat dalam memperoleh gelar
Sarjana Komputer



Disusun oleh:

NANDA RIO PRATAMA

22 05 3834

**PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TEKNOLOGI INFORMASI
UNIVERSITAS KRISTEN DUTA WACANA
TAHUN 2011**

PERNYATAAN KEASLIAN TUGAS AKHIR

Saya menyatakan dengan sesungguhnya bahwa tugas akhir dengan judul :

PENDETEKSIAN BAHASA PADA TEKS MENGGUNAKAN METODE N-GRAM

Yang saya kerjakan untuk melengkapi sebagian persyaratan menjadi Sarjana Komputer pada pendidikan sarjana Program Studi Teknik Informatika/Sistem Informasi, Fakultas Teknik Universitas Kristen Duta Wacana, bukan merupakan tiruan atau duplikasi dari skripsi kesarjanaan di lingkungan Universitas Kristen Duta Wacana maupun di Perguruan Tinggi atau instansi manapun, kecuali bagian yang sumber informasinya dicantumkan sebagaimana mestinya.

Jika dikemudian hari didapati bahwa tugas akhir ini adalah hasil plagiasi atau tiruan dari skripsi lain, saya bersedia menerima sanksi berupa pencabutan gelar kesarjanaan saya.



Yogyakarta, 29 Maret 2011

(NANDA RIO PRATAMA)
22 05 3834

HALAMAN PERSETUJUAN

Judul : PENDETEKSIAN BAHASA PADA TEKS
MENGUNAKAN METODE N-GRAM
Nama : NANDA RIO PRATAMA
NIM : 22 05 3834
Mata kuliah : Tugas Akhir
Kode : T12126
Semester : Genap
Tahun akademik : 2010/2011

Telah diperiksa dan disetujui
di Yogyakarta,
pada tanggal : 30 Maret 2011



Dosen Pembimbing I

(Lucia Dwi Krisnawati, S.S., MA.)

Dosen Pembimbing II

(Willy Sudiarto R, S.Kom., M.Cs.)

HALAMAN PENGESAHAN

PENDETEKSIAN BAHASA PADA TEKS MENGGUNAKAN METODE N-GRAM

Oleh : Nanda Rio Pratama / 22053834

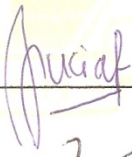


Dipertahankan di depan dewan Penguji Tugas Akhir/Skripsi
Program Studi Teknik Informatika Fakultas Teknologi Informasi
Universitas Kristen Duta Wacana – Yogyakarta
Dan dinyatakan diterima untuk memenuhi salah satu
syarat memperoleh gelar
Sarjana Komputer
Pada tanggal
11 April 2011

Yogyakarta, 28 April 2011

Mengesahkan,

Dewan Penguji :

1. Lucia Dwi Krisnawati, S.S., MA.
2. Willy Sudiarto R, S.Kom., M.Cs.
3. Ir. Sri Suwarno, M.Eng.

Dekan



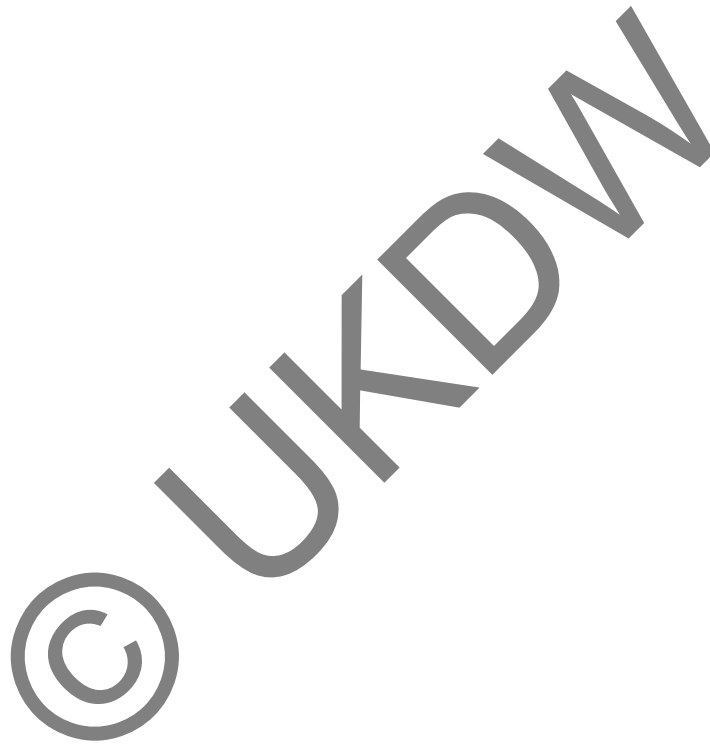

(Drs. Wimmie Handiwidjojo, MIT.)

Ketua Program Studi



(Nugroho Agus H, S.Si, M.Si.)

HALAMAN PERSEMBAHAN



PUJI SYUKUR KEPADA YESUS KRISTUS

SKRIPSI INI SAYA PERSEMBAHKAN KEPADA KELUARGA TERCINTA

DAN TEMAN SEPERJUANGAN

SEMOGA BERMANFAAT BAGI KITA SEMUA

UCAPAN TERIMA KASIH

Puji syukur penulis panjatkan kepada **Tuhan Yang Maha Esa** atas segala rahmat dan karunia serta pertolongan-Nya, sehingga penulis dapat menyelesaikan Tugas Akhir dengan judul Pendeteksian Bahasa Pada Teks Menggunakan Metode N-gram dengan baik dan tepat waktu.

Penulisan laporan ini merupakan kelengkapan dan pemenuhan dari salah satu syarat dalam memperoleh gelar Sarjana Komputer. Selain itu bertujuan melatih mahasiswa untuk dapat menghasilkan suatu karya yang dapat dipertanggungjawabkan secara ilmiah, sehingga dapat bermanfaat bagi penggunanya.

Dalam menyelesaikan program dan penyusunan laporan Tugas Akhir ini penulis telah banyak mendapatkan masukan dan bimbingan dari berbagai pihak untuk kelancaran penyelesaian penulisan Tugas Akhir ini. Untuk itu pada kesempatan ini penulis menyampaikan ucapan terimakasih kepada :

1. Ibu **Lucia Dwi Krisnawati, S.S., MA.**, selaku dosen pembimbing I yang telah banyak meluangkan waktunya memberikan pengarahan dan saran dari awal sampai terselesainya Tugas Akhir ini.
2. Bapak **Willy Sudiarto Raharjo, S.Kom., M.Cs.**, selaku dosen pembimbing II yang telah banyak memberi bimbingan dan petunjuk serta masukan-masukan dalam pembuatan Tugas Akhir ini.
3. Keluarga tercinta yang telah memberikan dukungan moral, dana, doa, saran dan kasih sayangnya yang berlimpah.
4. Teman-teman seperjuangan angkatan 2005 khususnya **Apul, Wahyu, Ragil, Popy, Chris** serta semua pihak yang tidak dapat penulis sebutkan satu persatu yang telah banyak memberi dukungan dan semangat dalam menyelesaikan tugas akhir ini.

Penulis menyadari bahwa program dan laporan Tugas Akhir ini masih jauh dari sempurna. Oleh karena itu, penulis sangat mengharapkan kritik dan saran yang

membangun dari pembaca, supaya suatu saat penulis dapat menghasilkan suatu karya yang lebih baik lagi.

Akhir kata penulis mohon maaf yang sebesar-besarnya apabila ada kesalahan selama penyusunan Tugas Akhir ini. Semoga Tugas Akhir ini dapat bermanfaat bagi kita semua.

Yogyakarta, Maret 2011

Penulis

© UKDWN

INTISARI

PENDETEKSIAN BAHASA PADA TEKS MENGUNAKAN METODE N-GRAM

Perkembangan teknologi saat ini semakin mempermudah manusia untuk saling bertukar informasi melalui Internet. Salah satu jenis informasi yang banyak ditemui adalah dalam bentuk teks. Untuk menyaring informasi mana yang relevan terhadap kebutuhan spesifik tiap orang, maka dibuat berbagai macam kategori-kategori yang salah satunya adalah berdasarkan bahasa. Oleh karena itu, dibutuhkan suatu sistem yang dapat melakukan pendeteksian bahasa secara otomatis menggunakan komputer agar pengkategorian dapat dilakukan secara lebih cepat dibandingkan dengan jika dilakukan secara manual oleh manusia.

Solusi dari permasalahan ini adalah dengan membuat suatu aplikasi yang dapat mendeteksi bahasa pada teks dengan menerapkan *statistical / computational approach* menggunakan metode *n-gram*.

Kesimpulan yang diperoleh penulis dari penelitian ini adalah, dengan menggunakan metode *n-gram* dalam melakukan pendeteksian bahasa pada teks akan menghasilkan suatu sistem yang dapat melakukan pendeteksian bahasa yang cukup akurat yaitu dengan rata-rata keberhasilan sebanyak 19,9125 dari 20 sampel dengan nilai parameter $n = 3$ dan maksrank atau jumlah *n-gram* sebanyak 400.

Kata Kunci : Pendeteksian bahasa, n-gram.

DAFTAR ISI

| | |
|--|--------------|
| HALAMAN JUDUL | |
| PERNYATAAN KEASLIAN TUGAS AKHIR..... | i |
| HALAMAN PERSETUJUAN | ii |
| HALAMAN PENGESAHAN..... | iii |
| HALAMAN PERSEMBAHAN | iv |
| UCAPAN TERIMA KASIH..... | v |
| INTISARI | vii |
| DAFTAR ISI | viii |
| DAFTAR GAMBAR..... | xi |
| DAFTAR TABEL | xii |
| | |
| BAB I PENDAHULUAN..... | 1 |
| 1.1 LATAR BELAKANG MASALAH..... | 1 |
| 1.2 RUMUSAN MASALAH | 2 |
| 1.3 BATASAN MASALAH..... | 3 |
| 1.4 TUJUAN PENULISAN | 3 |
| 1.5 METODE PENELITIAN | 3 |
| 1.6 SISTEMATIKA PENULISAN | 4 |
| | |
| BAB II LANDASAN TEORI | 5 |
| 2.1 TINJAUAN PUSTAKA | 5 |
| 2.2 LANDASAN TEORI | 5 |
| 2.2.1 Pendeteksian bahasa..... | 5 |
| 2.2.2 N-gram | 8 |
| 2.2.3 Penerapan N-gram dalam pendeteksian bahasa | 8 |
| 2.2.3.1 Pembentukan profil n-gram | 9 |
| 2.2.3.2 Proses membandingkan profil..... | 10 |

| | |
|--|-----------|
| BAB III PERANCANGAN SISTEM | 14 |
| 3.1 ANALISIS KEBUTUHAN | 14 |
| 3.1.1 Spesifikasi kemampuan sistem | 14 |
| 3.1.2 Spesifikasi kebutuhan sistem | 14 |
| 3.1.2.1 Kebutuhan perangkat lunak..... | 14 |
| 3.1.2.2 Kebutuhan perangkat keras minimal..... | 14 |
| 3.2 PERANCANGAN STRUKTUR DATA | 15 |
| 3.3 PERANCANGAN PROSES | 16 |
| 3.3.1 Proses pembuatan profil n-gram dari teks..... | 18 |
| 3.3.2 Proses pendeteksian bahasa pada teks..... | 21 |
| 3.4 PERANCANGAN ANTAR MUKA SISTEM | 23 |
| 3.4.1 Perancangan antarmuka pengaturan..... | 23 |
| 3.4.2 Perancangan antarmuka pembuatan sampel..... | 24 |
| 3.4.3 Perancangan antarmuka pendeteksian bahasa..... | 25 |
| 3.5 PERANCANGAN PENGUJIAN SISTEM..... | 25 |
| BAB IV IMPLEMENTASI DAN ANALISIS SISTEM | 26 |
| 4.1 IMPLEMENTASI SISTEM..... | 26 |
| 4.1.1 Antarmuka Sistem | 26 |
| 4.1.1.1 Antarmuka pendeteksian bahasa | 26 |
| 4.1.1.2 Antarmuka penambahan sampel | 27 |
| 4.1.1.3 Antarmuka pengaturan | 28 |
| 4.1.2 Format Masukan | 29 |
| 4.1.3 Bentuk keluaran | 30 |
| 4.1.4 Implementasi proses pendeteksian bahasa | 30 |
| 4.2 ANALISIS SISTEM..... | 33 |
| 4.2.1 Data sampel bahasa dan sampel teks | 33 |
| 4.2.2 Hasil Analisis Sistem | 34 |
| 4.2.2.1 Hasil Analisis n dan maksrank | 34 |

| | |
|--|-----------|
| 4.2.2.2 Hasil Analisis teks..... | 35 |
| 4.2.2.3 Hasil Pendeteksian Dengan Dokumen Training Yang Berbeda..... | 36 |
| 4.2.2.4 Hasil Pendeteksian Dengan dan Tanpa Unigram..... | 37 |
| BAB V KESIMPULAN Dan SARAN..... | 39 |
| 5.1 KESIMPULAN..... | 39 |
| 5.2 SARAN..... | 39 |

© UKDW

DAFTAR GAMBAR

| | |
|---|----|
| Gambar 2.1 Diagram proses identifikasi bahasa..... | 9 |
| Gambar 2.2 Gambar penghitungan jarak out of place | 11 |
| Gambar 3.1 Gambar struktur data sampel..... | 15 |
| Gambar 3.2 Gambar struktur data profil sampel..... | 16 |
| Gambar 3.3 Flowchart cara kerja sistem..... | 17 |
| Gambar 3.4 Flowchart Proses pembuatan profil n-gram dari teks I | 18 |
| Gambar 3.5 Flowchart proses pembuatan profil n-gram dari teks II | 19 |
| Gambar 3.6 Flowchart proses pendeteksian bahasa pada teks I | 22 |
| Gambar 3.7 Flowchart proses pendeteksian bahasa pada teks II | 23 |
| Gambar 3.8 Gambar antarmuka pengaturan | 23 |
| Gambar 3.9 Gambar antarmuka pembuatan sampel | 24 |
| Gambar 3.10 Gambar antarmuka pendeteksian bahasa | 25 |
| Gambar 4.1 Antarmuka Pendeteksian Bahasa | 27 |
| Gambar 4.2 Antarmuka penambahan sampel..... | 28 |
| Gambar 4.3 Antarmuka pengaturan | 29 |
| Gambar 4.4 Antarmuka proses masukan dari file | 29 |
| Gambar 4.5 Bentuk Keluaran..... | 30 |
| Gambar 4.6 Gambar rata-rata keberhasilan pendeteksian..... | 35 |
| Gambar 4.7 Grafik rata-rata hasil pendeteksian pada dokumen training yang berbeda | 37 |

DAFTAR TABEL

| | |
|--|----|
| Tabel 2.1 Tabel Pembuatan bi-gram pada teks..... | 10 |
| Tabel 2.2 Tabel Pembuatan profil bi-gram pada teks..... | 10 |
| Tabel 2.3 Tabel Profil dokumen A | 12 |
| Tabel 2.4 Tabel profil dokumen B..... | 12 |
| Tabel 2.5 Tabel proses perbandingan dokumen A dan dokumen B..... | 13 |

Tabel 4.1 Tabel hasil percobaan pada sistem34
Tabel 4.2 Tabel hasil kesalahan pendeteksian pada n 3 maksrank 400.....36

© UKDW

BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

Perkembangan teknologi saat ini semakin mempermudah manusia untuk saling bertukar informasi melalui Internet. Salah satu jenis informasi yang banyak ditemui adalah dalam bentuk teks. Untuk menyaring informasi mana yang relevan terhadap kebutuhan spesifik tiap orang, maka dibuat berbagai macam kategori-kategori yang salah satunya adalah berdasarkan bahasa. Oleh karena itu, dibutuhkan suatu sistem yang dapat melakukan pendeteksian bahasa secara otomatis menggunakan komputer agar pengkategorian dapat dilakukan secara lebih cepat dibandingkan dengan jika dilakukan secara manual oleh manusia.

Secara garis besar dalam melakukan pendeteksian bahasa dapat digunakan dua cara, yaitu *non-statistical / non-computational approach* dan secara *statistical / computational approach*. *Non stastical / non-computational approach* pada dasarnya adalah pendeteksian bahasa yang dilakukan secara manual oleh manusia dengan menggunakan suatu tabel kemunculan kata-kata (*table of frequent words*) dari suatu bahasa tertentu. Sedangkan pada *statistical / computational approach* pendeteksian bahasa dapat dilakukan menggunakan komputer. Salah satu metode yang populer digunakan adalah *common words and unique lettter combinations*. Ide dasar dari metode tersebut adalah dengan membandingkan kata-kata pada dokumen dengan suatu tabel berisi kata-kata kunci yang sering digunakan dalam setiap bahasa. Kelebihan dari metode ini adalah prosesnya yang sederhana dan cepat karena jika sudah menemukan satu atau beberapa kata kunci dari suatu bahasa maka proses pendeteksian sudah selesai. Tetapi sayangnya kelebihan tersebut diikuti dengan beberapa kelemahan, antara lain adalah pembuatan daftar kata kunci yang tidak mudah karena kata yang akan digunakan sebagai kata kunci pada suatu bahasa tertentu harus benar-benar merepresentasikan bahasa tersebut, berikutnya adalah jika

pada teks terdapat kesalahan penulisan maka kata yang seharusnya menjadi kata kunci tidak dapat digunakan, kelemahan yang lain adalah jika di dalam suatu teks mengandung banyak bahasa sekaligus seperti teks-teks dalam dunia kedokteran atau teknik yang banyak mengandung istilah-istilah yang tidak dapat diterjemahkan kedalam bahasa setempat sehingga kemungkinan terdapat beberapa kata kunci pada beberapa bahasa yang berbeda. Kelemahan-kelemahan tersebut dapat mengakibatkan hasil pendeteksian bahasa menjadi tidak tepat.

Pada tugas akhir ini, *statistical / computational approach* akan coba diimplementasikan untuk melakukan pendeteksian bahasa menggunakan metode *n-gram*. Pendeteksian bahasa dengan metode *n-gram* akan dilakukan dengan memecah kata-kata pada teks menjadi bagian yang lebih kecil dan disimpan dalam suatu profil berisi frekuensi dari tiap *gram* pada teks tersebut. Kemudian dilakukan penghitungan jarak antara profil-profil *n-gram* dari teks yang telah diketahui bahasanya dengan profil *n-gram* dari teks yang bahasanya akan dideteksi. Hasil penghitungan jarak tersebut nantinya akan menjadi dasar untuk mengambil keputusan mengenai bahasa apa yang digunakan pada teks. Dengan menggunakan metode *n-gram* diharapkan dapat mengatasi kelemahan-kelemahan pada metode *common words and unique letter combinations* sehingga menghasilkan suatu sistem yang lebih efisien dan dapat diandalkan dalam melakukan pendeteksian bahasa.

1.2 Rumusan Masalah

Berdasarkan uraian diatas maka masalah yang akan diteliti adalah sebagai berikut:

- Bagaimana mendeteksi bahasa pada teks menggunakan metode *n-gram*.
- Bagaimana akurasi dari sistem yang akan dibuat dalam melakukan pendeteksian bahasa.

1.3 Batasan Masalah

Mengingat banyaknya inputan yang bisa digunakan dalam penelitian ini maka penulis membatasi perumusan masalah sebagai berikut:

- Inputan pada sistem ini hanya berbentuk teks.
- Bahasa yang digunakan hanya berupa huruf latin.
- Tanda baca dan angka akan diabaikan.
- Banyaknya bahasa yang digunakan sebanyak 8 bahasa (Belanda, Indonesia, Inggris, Italia, Jerman, Prancis, Spanyol, Tagalog).

1.4 Tujuan Penulisan

Melalui penelitian ini tujuan yang ingin dicapai penulis adalah: Membuat sistem pendeteksi bahasa menggunakan metode *n-gram*.

1.5 Metode Penelitian

Metodologi yang digunakan dalam penelitian ini adalah sebagai berikut:

- Mempelajari sumber pustaka yang berkaitan dengan *n-gram* dan pendeteksian bahasa baik berupa buku maupun sumber on-line dari Internet.
- Analisis kebutuhan dan kondisi dari aplikasi yang akan dibangun baik pada tingkat perangkat lunak maupun perangkat keras.
- Perancangan sistem dan aplikasi yang akan dibangun.
- Implementasi hasil perancangan dalam kedalam bahasa pemrograman (*coding*).
- Pengujian dan analisis terhadap aplikasi yang telah dibangun.
- Penarikan kesimpulan.

1.6 Sistematika Penulisan

Sistematika penulisan laporan tugas akhir ini dibagi menjadi beberapa bab sebagai berikut:

Bab 1 berupa PENDAHULUAN, berisi latar belakang masalah yang akan diteliti dan rencana penelitian yang akan dilakukan. Bab 2 merupakan TINJAUAN PUSTAKA yang memuat uraian dari konsep-konsep atau teori-teori yang digunakan sebagai dasar pembuatan tugas akhir ini. Bab 3 merupakan PERANCANGAN SISTEM, yang berisi tahapan dalam perancangan dan pembangunan sistem, termasuk aliran data dan rancangan antarmuka form masukan (input) dan form hasil (output) beserta kegunaannya. Bab 4 merupakan IMPLEMENTASI dan ANALISIS SISTEM, membahas tentang implementasi perancangan sistem pada bab 3 beserta analisis dan hasil capture dari sistem yang dibuat. Bab 5 merupakan KESIMPULAN dan SARAN, berisi kesimpulan dari hasil penelitian yang dilakukan serta memberikan saran untuk pengembangan penelitian yang telah dilakukan.



BAB V

KESIMPULAN Dan SARAN

5.1 Kesimpulan

Dari hasil penelitian yang dilakukan maka dapat disimpulkan bahwa tidak semua teks dapat dideteksi oleh sistem secara benar, beberapa kriteria teks yang dapat dideteksi oleh sistem adalah, pertama hanya terdiri dari satu bahasa, yang kedua jumlah kata pada teks paling tidak lebih dari sepuluh kata, dan ketiga teks harus berupa huruf latin. Selain itu, dari hasil percobaan ditemukan bahwa program dapat memberikan hasil pendeteksian yang paling akurat dari sampel-sampel yang digunakan pada 3 untuk n dan untuk maksrank atau jumlah n -gram sebanyak 400 dengan rata-rata keberhasilan sebanyak 19,9125 dari 20 sampel atau sebesar 99,5625%.

5.2 Saran

Selain kesimpulan diatas, penulis juga mempunyai beberapa saran untuk pengembangan sistem antara lain adalah pengembangan sistem untuk mendeteksi bahasa-bahasa pada sebuah teks yang terdiri dari beberapa bahasa sekaligus, selain itu juga dapat dilakukan pengembangan untuk melakukan pendeteksian bahasa pada teks pendek atau bahkan pada teks yang hanya terdiri dari satu kata, dan yang terakhir adalah dapat dilakukan pengembangan untuk mendeteksi teks yang bukan merupakan huruf latin.

DAFTAR PUSTAKA

- Ahmed, B., Cha, Sung-Hyuk. and Tappert, C. 2004. Language Identification from Text Using N-gram Based Cumulative Frequency Addition. *Proceedings of Student/Faculty Research Day, CSIS, Pace University*.
- Baldwin, T. 2009. Data on the Web, Language Identification. Power Point Slides. Diakses 10 Juni 2010 dari <http://www.cs.mu.oz.au/352/lectures/handout03b.pdf>
- Cavnar, W. 1994. Using an n-gram-based document representation with a vector processing retrieval model. *In Proceedings of the Third Text Retrieval Conference (TREC-3)*.
- Cavnar, W., and J.M. Trenkle. 1994. N-gram based text categorization. *In Proceedings of Symposium on Document Analysis and Information Retrieval*.
- Dunning, T. 1994. Statistical Identification of Language. *Technical report CRL*. Computing Research Lab, New Mexico State University.
- Gold, E. M. 1967. Language identification in the limit. *Journal Information and Control*. Vol 10, 447-474.
- Olvecky, T. 2005. N-Gram Based Statistics Aimed at Language Identification. *M. Bieliková (Ed.), IIT.SRC 2005*, 1-7.
- Peng, F. 2003. Language Independent Text Learning with Statistical n Gram Language Models. Thesis, University of Waterloo, Ontario, Canada.